(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7362929号
(P7362929)

(45)発行日　令和5年10月17日(2023.10.17)

(24)登録日　令和5年10月6日(2023.10.6)

(51)国際特許分類

Ｇ１０Ｌ　13/10　(2013.01)	Ｇ１０Ｌ	13/10	１１３Ｚ
Ｇ１０Ｌ　25/30　(2013.01)	Ｇ１０Ｌ	25/30	

請求項の数　28　（全27頁）

(21)出願番号　特願2022-534694(P2022-534694)
(86)(22)出願日　令和1年12月10日(2019.12.10)
(65)公表番号　特表2023-505670(P2023-505670A)
(43)公表日　令和5年2月10日(2023.2.10)
(86)国際出願番号　PCT/US2019/065566
(87)国際公開番号　WO2021/118543
(87)国際公開日　令和3年6月17日(2021.6.17)
審査請求日　令和4年7月20日(2022.7.20)

(73)特許権者　502208397
グーグル エルエルシー
Ｇｏｏｇｌｅ ＬＬＣ
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン ビュー アンフィシアター パークウェイ １６００
１６００ Ａｍｐｈｉｔｈｅａｔｒｅ Ｐａｒｋｗａｙ ９４０４３ Ｍｏｕｎｔａｉｎ Ｖｉｅｗ， ＣＡ Ｕ．Ｓ．Ａ．
(74)代理人　100142907
弁理士　本田 淳
(72)発明者　クラーク、ロバート
アメリカ合衆国 ９４０４３ カリフォルニア州 マウンテン ビュー アンフィシアター パークウェイ １６００

最終頁に続く

(54)【発明の名称】　アテンションベースのクロックワーク階層型変分エンコーダ

(57)【特許請求の範囲】

【請求項１】

方法（４００）であって、

データ処理ハードウェア（１１２）において、少なくとも１つの単語（２４０）を有するテキスト発話（３１０）を受信することであって、各単語（２４０）は少なくとも１つの音節（２３０）を有し、各音節（２３０）は少なくとも１つの音素（２２０）を有する、テキスト発話（３１０）を受信すること、

前記データ処理ハードウェア（１１２）によって、前記テキスト発話（３１０）のための発話埋め込み（２０４）であって、意図された韻律を表す前記発話埋め込み（２０４）を選択すること、

選択された発話埋め込み（２０４）を用いて、各音節（２３０）について、

前記データ処理ハードウェア（１１２）によって、前記音節（２３０）の各音素（２２０）の言語的特徴（２２２）に対するアテンション機構（３４０）によるアテンションに基づいて、前記音節（２３０）の韻律音節埋め込み（２３２，２３４）を復号化することにより、前記音節（２３０）の持続時間（２３８）を予測すること、

前記データ処理ハードウェア（１１２）によって、前記音節（２３０）の予測された持続時間（２３８）に基づいて、複数の固定長予測フレーム（２６０）を生成すること

を含む、方法（４００）。

【請求項２】

前記データ処理ハードウェア（１１２）によって、前記音節（２３０）の前記予測され

た持続時間（２３８）に基づいて前記音節（２３０）のピッチ輪郭（Ｆ０）を予測することをさらに含み、
前記複数の固定長予測フレーム（２６０）は、固定長予測ピッチフレーム（２６０，２６０Ｆ０）を含み、各固定長予測ピッチフレーム（２６０Ｆ０）は、前記音節（２３０）の予測されたピッチ輪郭（Ｆ０）の一部を表す、請求項１に記載の方法（４００）。
【請求項３】
前記選択された発話埋め込み（２０４）を用いて、各音節（２３０）について、
前記データ処理ハードウェア（１１２）によって、前記音節（２３０）の前記予測された持続時間（２３８）に基づいて、各音節（２３０）のエネルギー輪郭（Ｃ０）を予測すること、
前記データ処理ハードウェア（１１２）によって、対応する音節（２３０）の前記予測された持続時間（２３８）に基づいて、複数の固定長予測エネルギーフレーム（２６０，２６０Ｃ０）を生成すること
をさらに含み、各固定長エネルギーフレーム（２６０Ｃ０）は、前記対応する音節（２３０）の予測されたエネルギー輪郭（Ｃ０）を表す、請求項１または２に記載の方法（４００）。
【請求項４】
前記複数の固定長予測フレーム（２６０）は、前記音節（２３０）の固定長予測スペクトルフレーム（２６０，２６０Ｍ０）を含む、請求項１～３のいずれか一項に記載の方法（４００）。
【請求項５】
前記テキスト発話（３１０）の階層的言語構造（２００）を表すネットワークは、
前記テキスト発話（３１０）の各単語（２４０）を含む第１レベルと、
前記テキスト発話（３１０）の各音節（２３０）を含む第２レベルと、
前記テキスト発話（３１０）の各音節（２３０）の各固定長予測フレーム（２６０）を含む第３レベルと
を含む、請求項１～４のいずれか一項に記載の方法（４００）。
【請求項６】
前記階層的言語構造（２００）を表す前記ネットワークの前記第１レベルは、前記テキスト発話（３１０）の各単語（２４０）を表す長短期記憶（ＬＳＴＭ）処理ブロックを含み、
前記階層的言語構造（２００）を表す前記ネットワークの前記第２レベルは、前記テキスト発話（３１０）の各音節（２３０）を表すＬＳＴＭ処理ブロックを含み、前記第２レベルの前記ＬＳＴＭ処理ブロックは、前記第１レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ前記第１レベルのＬＳＴＭ処理ブロックよりも高速にクロックし、
前記階層的言語構造（２００）を表す前記ネットワークの前記第３レベルは、各固定長予測フレーム（２６０）を表すＬＳＴＭ処理ブロックを含み、前記第３レベルの前記ＬＳＴＭ処理ブロックは、前記第２レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ前記第２レベルのＬＳＴＭ処理ブロックよりも高速にクロックする、請求項５に記載の方法（４００）。
【請求項７】
前記音節（２３０）の前記持続時間（２３８）を予測することは、
前記音節（２３０）に関連付けられた各音素（２２０）について、
対応する音素（２２０）の１つまたは複数の言語的特徴（２２２）を符号化すること、
符号化された１つまたは複数の言語的特徴（２２２）を前記アテンション機構（３４０）に入力すること、
前記アテンション機構（３４０）の前記アテンションを前記韻律音節埋め込み（２３２，２３４）に適用すること
を含む、請求項１～６のいずれか一項に記載の方法（４００）。
【請求項８】

前記韻律音節埋め込み（２３２，２３４）は、前記発話埋め込み（２０４）に対応するフレーム（２１０）に基づく第１の音節埋め込み（２３２）と、前記発話埋め込み（２０４）の１つまたは複数の音素（２２０）に関連付けられた音素言語的特徴（２２２）に基づく第２の音節埋め込み（２３４）とを含む、請求項１～７のいずれか一項に記載の方法（４００）。

【請求項９】

前記データ処理ハードウェア（１１２）によって、複数の基準オーディオ信号（２０２）を含むトレーニングデータを受信することであって、各基準オーディオ信号（２０２）は、人間音声の音声発話を含み、かつ対応する韻律を有する、トレーニングデータを受信すること、

前記データ処理ハードウェア（１１２）によって、各基準オーディオ信号（２０２）を、前記基準オーディオ信号（２０２）の前記対応する韻律を表す対応する固定長発話埋め込み（２０４）に符号化することにより、韻律モデル（３００）のためのディープニューラルネットワーク（２００）をトレーニングすること

をさらに含む、請求項１～８のいずれか一項に記載の方法（４００）。

【請求項１０】

前記データ処理ハードウェア（１１２）によって、フレームベースの音節埋め込み（２３２）および単音特徴ベースの音節埋め込み（２３４）で、複数の言語ユニット（２２０，２３０，２４０，２５０）の言語的特徴（２２２，２３６，２４２，２５２）を符号化することにより、前記選択された発話埋め込み（２０４）を生成することをさらに含む、請求項９に記載の方法（４００）。

【請求項１１】

前記発話埋め込み（２０４）は、固定長の数値ベクトルを含む、請求項１～１０のいずれか一項に記載の方法（４００）。

【請求項１２】

前記アテンション機構（３４０）の前記アテンションは、位置ベースのアテンションを含む、請求項１～１１のいずれか一項に記載の方法（４００）。

【請求項１３】

前記位置ベースのアテンションは、単調にシフトする、位置に敏感なアテンションを含み、前記単調にシフトする、位置に敏感なアテンションは、それぞれの音節（２３０）の音素情報のウィンドウによって定義される、請求項１２に記載の方法（４００）。

【請求項１４】

前記アテンション機構（３４０）は、トランスフォーマーを含む、請求項１～１１のいずれか一項に記載の方法（４００）。

【請求項１５】

データ処理ハードウェア（１１２）と、

前記データ処理ハードウェア（１１２）と通信するメモリハードウェア（１１４）と

を備えるシステム（１００）であって、前記メモリハードウェア（１１４）は、前記データ処理ハードウェア（１１２）上で実行されると前記データ処理ハードウェア（１１２）に動作を実行させる命令を格納しており、前記動作は、

少なくとも１つの単語（２４０）を有するテキスト発話（３１０）を受信することであって、各単語（２４０）は少なくとも１つの音節（２３０）を有し、各音節（２３０）は少なくとも１つの音素（２２０）を有する、テキスト発話（３１０）を受信すること、

前記テキスト発話（３１０）のための発話埋め込み（２０４）であって、意図された韻律を表す前記発話埋め込み（２０４）を選択すること、

選択された発話埋め込み（２０４）を用いて、各音節（２３０）について、

前記音節（２３０）の各音素（２２０）の言語的特徴（２２２）に対するアテンション機構（３４０）によるアテンションに基づいて、前記音節（２３０）の韻律音節埋め込み（２３２，２３４）を復号化することにより、前記音節（２３０）の持続時間（２３８）を予測すること、

前記音節（２３０）の予測された持続時間（２３８）に基づいて、複数の固定長予測フレーム（２６０）を生成すること
を含む、システム（１００）。
【請求項１６】
前記動作は、
前記音節（２３０）の前記予測された持続時間（２３８）に基づいて前記音節（２３０）のピッチ輪郭（Ｆ０）を予測すること
をさらに含み、
前記複数の固定長予測フレーム（２６０）は、固定長予測ピッチフレーム（２６０，２６０Ｆ０）を含み、各固定長予測ピッチフレーム（２６０Ｆ０）は、前記音節（２３０）の予測されたピッチ輪郭（Ｆ０）の一部を表す、請求項１５に記載のシステム（１００）。
【請求項１７】
前記動作は、前記選択された発話埋め込み（２０４）を用いて、各音節（２３０）について、
前記音節（２３０）の前記予測された持続時間（２３８）に基づいて、各音節（２３０）のエネルギー輪郭（Ｃ０）を予測すること、
対応する音節（２３０）の前記予測された持続時間（２３８）に基づいて、複数の固定長予測エネルギーフレーム（２６０，２６０Ｃ０）を生成すること
をさらに含み、各固定長エネルギーフレーム（２６０Ｃ０）は、前記対応する音節（２３０）の予測されたエネルギー輪郭（Ｃ０）を表す、請求項１５または１６に記載のシステム（１００）。
【請求項１８】
前記複数の固定長予測フレーム（２６０）は、前記音節（２３０）の固定長予測スペクトルフレーム（２６０，２６０Ｍ０）を含む、請求項１５～１７のいずれか一項に記載のシステム（１００）。
【請求項１９】
前記テキスト発話（３１０）の階層的言語構造（２００）を表すネットワークは、
前記テキスト発話（３１０）の各単語（２４０）を含む第１レベルと、
前記テキスト発話（３１０）の各音節（２３０）を含む第２レベルと、
前記テキスト発話（３１０）の各音節（２３０）の各固定長予測フレーム（２６０）を含む第３レベルと
を含む、請求項１５～１８のいずれか一項に記載のシステム（１００）。
【請求項２０】
前記階層的言語構造（２００）を表す前記ネットワークの前記第１レベルは、前記テキスト発話（３１０）の各単語（２４０）を表す長短期記憶（ＬＳＴＭ）処理ブロックを含み、
前記階層的言語構造（２００）を表す前記ネットワークの前記第２レベルは、前記テキスト発話（３１０）の各音節（２３０）を表すＬＳＴＭ処理ブロックを含み、前記第２レベルの前記ＬＳＴＭ処理ブロックは、前記第１レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ前記第１レベルのＬＳＴＭ処理ブロックよりも高速にクロックし、
前記階層的言語構造（２００）を表す前記ネットワークの前記第３レベルは、各固定長予測フレーム（２６０）を表すＬＳＴＭ処理ブロックを含み、前記第３レベルの前記ＬＳＴＭ処理ブロックは、前記第２レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ前記第２レベルのＬＳＴＭ処理ブロックよりも高速にクロックする、請求項１９に記載のシステム（１００）。
【請求項２１】
前記音節（２３０）の前記持続時間（２３８）を予測することは、
前記音節（２３０）に関連付けられた各音素（２２０）について、
対応する音素（２２０）の１つまたは複数の言語的特徴（２２２）を符号化すること、
符号化された１つまたは複数の言語的特徴（２２２）を前記アテンション機構（３４

０）に入力すること、
　前記アテンション機構（３４０）の前記アテンションを前記韻律音節埋め込み（２３２，２３４）に適用すること
を含む、請求項１５～２０のいずれか一項に記載のシステム（１００）。
【請求項２２】
前記韻律音節埋め込み（２３２，２３４）は、前記発話埋め込み（２０４）に対応するフレーム（２１０）に基づく第１の音節埋め込み（２３２）と、前記発話埋め込み（２０４）の１つまたは複数の音素（２２０）に関連付けられた音素言語的特徴（２２２）に基づく第２の音節埋め込み（２３４）とを含む、請求項１５～２１のいずれか一項に記載のシステム（１００）。
【請求項２３】
前記動作は、
複数の基準オーディオ信号（２０２）を含むトレーニングデータを受信することであって、各基準オーディオ信号（２０２）は、人間音声の音声発話を含み、かつ対応する韻律を有する、トレーニングデータを受信すること、
各基準オーディオ信号（２０２）を、前記基準オーディオ信号（２０２）の前記対応する韻律を表す対応する固定長発話埋め込み（２０４）に符号化することにより、韻律モデル（３００）のためのディープニューラルネットワーク（２００）をトレーニングすること
をさらに含む、請求項１５～２２のいずれか一項に記載のシステム（１００）。
【請求項２４】
前記動作は、フレームベースの音節埋め込み（２３２）および単音特徴ベースの音節埋め込み（２３４）で、複数の言語ユニット（２２０，２３０，２４０，２５０）の言語的特徴（２２２，２３６，２４２，２５２）を符号化することにより、前記選択された発話埋め込み（２０４）を生成することをさらに含む、請求項２３に記載のシステム（１００）。
【請求項２５】
前記発話埋め込み（２０４）は、固定長の数値ベクトルを含む、請求項１５～２４のいずれか一項に記載のシステム（１００）。
【請求項２６】
前記アテンション機構（３４０）の前記アテンションは、位置ベースのアテンションを含む、請求項１５～２５のいずれか一項に記載のシステム（１００）。
【請求項２７】
前記位置ベースのアテンションは、単調にシフトする、位置に敏感なアテンションを含み、前記単調にシフトする、位置に敏感なアテンションは、それぞれの音節（２３０）の音素情報のウィンドウによって定義される、請求項２６に記載のシステム（１００）。
【請求項２８】
前記アテンション機構（３４０）は、トランスフォーマーを含む、請求項１５～２５のいずれか一項に記載のシステム（１００）。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、アテンションベースのクロックワーク階層型変分エンコーダに関する。
【背景技術】
【０００２】
音声合成システムは、テキスト音声変換（ｔｅｘｔ－ｔｏ－ｓｐｅｅｃｈ，ＴＴＳ）モデルを使用して、テキスト入力から音声を生成する。生成／合成された音声は、意図された韻律（表現力）を備えた人間の音声のように聞こえながら（自然さ）、メッセージを正確に伝える（分かりやすさ）必要がある。従来の波形接続型（ｃｏｎｃａｔｅｎａｔｉｖｅ）およびパラメトリック合成モデルは分かりやすい音声を提供でき、音声のニューラルモデリングの最近の進歩により、合成音声の自然さが大幅に向上したが、既存のＴＴＳモ

デルのほとんどは韻律のモデリングに効果がないため、重要なアプリケーションで使用される合成音声に表現力が不足する原因となっている。例えば、会話アシスタントおよび長文リーダー（ｌｏｎｇ－ｆｏｒｍ　ｒｅａｄｅｒ）などのアプリケーションでは、イントネーション、強勢、リズムおよびスタイルのような、テキスト入力では伝達されない韻律的な特徴を入力することで、リアルな音声を生成することが望ましい。例えば、単純なステートメントは、ステートメントが質問であるか、質問への回答であるか、ステートメントに不確実性があるか、または入力テキストによって指定されていない環境またはコンテキストに関するその他の意味を伝えるかどうかに応じて、さまざまな方法で話すことができる。

【発明の概要】

【０００３】

本開示の一態様は、アテンション（ａｔｔｅｎｔｉｏｎ）ベースのクロックワーク階層型変分エンコーダ（ｃｌｏｃｋｗｏｒｋ　ｈｉｅｒａｒｃｈｉｃａｌ　ｖａｒｉａｔｉｏｎａｌ　ｅｎｃｏｄｅｒ）のための方法を提供する。方法は、データ処理ハードウェアにおいて、少なくとも１つの単語を有するテキスト発話を受信することを含み、各単語は少なくとも１つの音節を有し、各音節は少なくとも１つの音素を有している。方法は、データ処理ハードウェアによって、テキスト発話のための発話埋め込みを選択することも含む。発話埋め込みは、意図された韻律を表す。方法は、選択された発話埋め込みを用いて、各音節について、データ処理ハードウェアによって、音節の各音素の言語的特徴に対するアテンション機構（ａｔｔｅｎｔｉｏｎ　ｍｅｃｈａｎｉｓｍ）によるアテンションに基づいて、音節の韻律音節埋め込みを復号化することにより、音節の持続時間（ｄｕｒａｔｉｏｎ）を予測することを含む。方法は、選択された発話埋め込みを用いて、各音節について、データ処理ハードウェアによって、音節の予測された持続時間に基づいて、複数の固定長予測フレームを生成することを含む。

【０００４】

本開示の実装形態は、以下の任意選択の特徴のうちの１つまたは複数を含んでいてよい。いくつかの実装形態では、方法は、データ処理ハードウェアによって、音節の予測された持続時間に基づいて音節のピッチ輪郭を予測することを含む。この実装形態において、複数の固定長予測フレームが固定長予測ピッチフレームを含む場合、各固定長予測ピッチフレームは、音節の予測されたピッチ輪郭の一部を表している。

【０００５】

いくつかの例において、方法は、選択された発話埋め込みを用いて、各音節について、データ処理ハードウェアによって、音節の予測された持続時間に基づいて、各音節のエネルギー輪郭を予測することを含む。この例では、方法は、データ処理ハードウェアによって、対応する音節の予測された持続時間に基づいて、複数の固定長予測エネルギーフレームを生成することも含み、各固定長エネルギーフレームは、対応する音節の予測されたエネルギー輪郭を表す。複数の固定長予測フレームは、音節の固定長予測スペクトルフレームを含み得る。

【０００６】

いくつかの構成では、テキスト発話の階層的言語構造を表すネットワークは、テキスト発話の各単語を含む第１レベルと、テキスト発話の各音節を含む第２レベルと、テキスト発話の各音節の各固定長予測フレームを含む第３レベルとを含む。ここで、階層的言語構造を表すネットワークの第１レベルは、テキスト発話の各単語を表す長短期記憶（ＬＳＴＭ）処理ブロックを含み得る。階層的言語構造を表すネットワークの第２レベルは、テキスト発話の各音節を表すＬＳＴＭ処理ブロックを含んでいてよく、第２レベルのＬＳＴＭ処理ブロックは、第１レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ第１レベルのＬＳＴＭ処理ブロックよりも高速にクロックする。階層的言語構造を表すネットワークの第３レベルは、各固定長予測フレームを表すＬＳＴＭ処理ブロックを含んでいてよく、第３レベルのＬＳＴＭ処理ブロックは、第２レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ第２レベルのＬＳＴＭ処理ブロックよりも高速にクロックする。

【０００７】
いくつかの構成では、音節の長さを予測することは、音節に関連付けられた各音素について、対応する音素の１つまたは複数の言語的特徴を符号化すること、符号化された１つまたは複数の言語的特徴をアテンション機構に入力すること、アテンション機構のアテンションを韻律音節埋め込みに適用する（ａｐｐｌｙｉｎｇ）ことを含む。韻律音節埋め込みは、発話埋め込みに対応するフレームに基づく第１の音節埋め込みと、発話埋め込みの１つまたは複数の音素に関連付けられた音素言語的特徴に基づく第２の音節埋め込みとを含み得る。
【０００８】
いくつかの例では、方法は、データ処理ハードウェアによって、複数の基準オーディオ信号を含むトレーニングデータを受信すること、データ処理ハードウェアによって、各基準オーディオ信号の対応する韻律を表す対応する固定長発話埋め込みに符号化することにより、韻律モデルのためのディープニューラルネットワークをトレーニングすることを含む。この例では、各基準オーディオ信号は、人間音声の音声発話を含み、かつ対応する韻律を有している。ここで、方法は、データ処理ハードウェアによって、フレームベースの音節埋め込みおよび単音（ｐｈｏｎｅ）特徴ベースの音節埋め込みで、複数の言語ユニットの言語的特徴を符号化することにより、選択された発話埋め込みを生成することを含み得る。発話埋め込みは、固定長の数値ベクトルを含んでいてよい。
【０００９】
いくつかの実装形態では、アテンション機構のアテンションは、位置ベースのアテンションを含む。位置ベースのアテンションは、単調にシフトする、位置に敏感な（ｌｏｃａｔｉｏｎ　ｓｅｎｓｉｔｉｖｅ）アテンションを含み、単調にシフトする、位置に敏感なアテンションは、それぞれの音節の音素情報のウィンドウによって定義される。アテンション機構は、トランスフォーマー（ｔｒａｎｓｆｏｒｍｅｒ）を含み得る。
【００１０】
本開示の別の態様は、アテンションベースのクロックワーク階層型変分エンコーダのためのシステムを提供する。システムは、データ処理ハードウェアおよびデータ処理ハードウェアと通信するメモリハードウェアを含む。メモリハードウェアは、データ処理ハードウェア上で実行されると、データ処理ハードウェアに動作を実行させる命令を格納している。動作は、少なくとも１つの単語を有するテキスト発話を受信することを含み、各単語は少なくとも１つの音節を有し、各音節は少なくとも１つの音素を有している。動作は、テキスト発話のための発話埋め込みを選択することも含み、発話埋め込みは、意図された韻律を表す。動作は、選択された発話埋め込みを用いて、各音節について、音節の各音素の言語的特徴に対するアテンション機構によるアテンションに基づいて、音節の韻律音節埋め込みを復号化することにより、音節の持続時間を予測することをさらに含む。動作は、選択された発話埋め込みを用いて、各音節について、音節の予測された持続時間に基づいて、複数の固定長予測フレームを生成することも含む。
【００１１】
この態様は、以下の任意選択の特徴のうちの１つまたは複数を含んでいてよい。いくつかの構成では、動作は、音節の予測された持続時間に基づいて音節のピッチ輪郭を予測することを含み、複数の固定長予測フレームが固定長予測ピッチフレームを含む場合、各固定長予測ピッチフレームは、音節の予測されたピッチ輪郭の一部を表している。動作は、選択された発話埋め込みを用いて、各音節について、音節の予測された持続時間に基づいて、各音節のエネルギー輪郭を予測すること、対応する音節の予測された持続時間に基づいて、複数の固定長予測エネルギーフレームを生成することを含んでいてもよく、各固定長エネルギーフレームは、対応する音節の予測されたエネルギー輪郭を表す。複数の固定長予測フレームは、音節の固定長予測スペクトルフレームを含み得る。
【００１２】
いくつかの例では、テキスト発話の階層的言語構造を表すネットワークは、テキスト発話の各単語を含む第１レベルと、テキスト発話の各音節を含む第２レベルと、テキスト発

話の各音節の各固定長予測フレームを含む第３レベルとを含む。ここで、階層的言語構造を表すネットワークの第１レベルは、テキスト発話の各単語を表す長短期記憶（ＬＳＴＭ）処理ブロックを含み得る。階層的言語構造を表すネットワークの第２レベルは、テキスト発話の各音節を表すＬＳＴＭ処理ブロックを含んでいてよく、第２レベルのＬＳＴＭ処理ブロックは、第１レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ第１レベルのＬＳＴＭ処理ブロックよりも高速にクロックする。階層的言語構造を表すネットワークの第３レベルは、各固定長予測フレームを表すＬＳＴＭ処理ブロックを含んでいてよく、第３レベルのＬＳＴＭ処理ブロックは、第２レベルのＬＳＴＭ処理ブロックに対して相対的に、かつ第２レベルのＬＳＴＭ処理ブロックよりも高速にクロックする。

【００１３】

いくつかの実装形態では、音節の長さを予測することは、音節に関連付けられた各音素について、対応する音素の１つまたは複数の言語的特徴を符号化すること、符号化された１つまたは複数の言語的特徴をアテンション機構に入力すること、アテンション機構のアテンションを韻律音節埋め込みに適用することを含む。韻律音節埋め込みは、発話埋め込みに対応するフレームに基づく第１の音節埋め込みと、発話埋め込みの１つまたは複数の音素に関連付けられた音素言語的特徴に基づく第２の音節埋め込みとを含み得る。

【００１４】

いくつかの構成では、動作は、複数の基準オーディオ信号を含むトレーニングデータを受信することを含み、各基準オーディオ信号は、人間音声の音声発話を含み、かつ対応する韻律を有する。この構成では、動作は、各基準オーディオ信号を、基準オーディオ信号の対応する韻律を表す対応する固定長発話埋め込みに符号化することにより、韻律モデルのためのディープニューラルネットワークをトレーニングすることも含む。ここで、動作は、フレームベースの音節埋め込みおよび単音特徴ベースの音節埋め込みで、複数の言語ユニットの言語的特徴を符号化することにより、選択された発話埋め込みを生成することを含み得る。発話埋め込みは、固定長の数値ベクトルを含んでいてよい。

【００１５】

いくつかの例では、アテンション機構のアテンションは、位置ベースのアテンションを含む。ここで、位置ベースのアテンションは、単調にシフトする、位置に敏感なアテンションを含み、単調にシフトする、位置に敏感なアテンションは、それぞれの音節の音素情報のウィンドウによって定義される。アテンション機構は、トランスフォーマーを含み得る。

【００１６】

本開示の１つまたは複数の実施形態の詳細は、添付の図面および以下の説明に記載されている。他の態様、特徴、および利点は、説明および図面、ならびに特許請求の範囲から明らかになるであろう。

【図面の簡単な説明】

【００１７】

【図１】図１は、テキスト発話の韻律表現を予測する際に使用する制御可能な韻律モデルを提供するためにディープニューラルネットワークをトレーニングするための例示的なシステムの概略図である。

【図２Ａ】図２Ａは、基準オーディオ信号の韻律を固定長の発話埋め込みに符号化するための階層的言語構造の概略図である。

【図２Ｂ】図２Ｂは、テキスト発話の韻律表現を予測するために発話埋め込みを使用する階層的言語構造の概略図である。

【図２Ｃ】図２Ｃは、テキスト発話の韻律表現を予測するために発話埋め込みを使用する階層的言語構造の概略図である。

【図３Ａ】図３Ａは、テキスト発話の音節特性を予測するための例示的なオートエンコーダの概略図である。

【図３Ｂ】図３Ｂは、テキスト発話の音節特性を予測するための例示的なオートエンコーダの概略図である。

【図３Ｃ】図３Ｃは、テキスト発話の音節特性を予測するための例示的なオートエンコーダの概略図である。
【図３Ｄ】図３Ｄは、符号化された音素状態を形成する単音レベルの特徴にアテンションを与えるように構成された例示的なアテンション機構の概略図である。
【図４】図４は、受信されたテキスト発話の表現を予測する方法のための動作の例示的な配置のフローチャートである。
【図５】図５は、本明細書で説明されるシステムおよび方法を実装するために使用され得る例示的なコンピューティングデバイスの概略図である。
【発明を実施するための形態】
【００１８】
さまざまな図面の同様の参照記号は、同様の要素を示す。
音声合成システムでしばしば使用されるテキスト音声変換（Ｔｅｘｔ－ｔｏ－Ｓｐｅｅｃｈ，ＴＴＳ）モデルは、概して、ランタイムにおいて、いかなる基準音響表現もなしでテキスト入力のみを与えられ、本物らしく聞こえる合成音声を生成するために、テキスト入力によって提供されない多くの言語的ファクターを帰属させる必要がある。これらの言語的ファクターのサブセットは、まとめて韻律と呼ばれ、イントネーション（ピッチの変化）、強勢（強勢のある音節対無強勢の音節）、音の持続時間、ラウドネス、トーン、リズム、および音声のスタイルを含んでいてよい。韻律は、音声の感情状態、音声の形式（例えば、ステートメント、質問、コマンドなど）、音声の皮肉または嫌味の存在、音声の認識における不確実性、または入力テキストの文法または語彙選択によって符号化されることが不可能な他の言語要素を示し得る。したがって、大きな韻律変化に関連付けられた特定のテキスト入力は、ピッチおよび発話継続時間の局所的な変化を伴う合成音声を生成して、異なるセマンティック意味を伝えることができ、また、全体的なピッチ軌道のグローバルな変化を伴う合成音声を生成して、異なる気分および感情を伝えることもできる。
【００１９】
ニューラルネットワークモデルは、テキスト入力では提供されない韻律に対応する言語的ファクターを予測することにより、音声を確実に合成する可能性を提供する。その結果、オーディオブックのナレーション、ニュースリーダー、ボイスデザインソフトウェア、および会話アシスタントなどの複数のアプリケーションが、単調には響かない本物らしく聞こえる合成音声を生成することができる。本明細書の実装形態は、音声発話に対応する基準オーディオ信号を音声発話の韻律を表す発話埋め込みに符号化するためのエンコーダ部分と、発話埋め込みを復号化して、音節の持続時間と、各音節のピッチおよびエネルギー輪郭とを予測するデコーダ部分とを有する変分オートエンコーダ（ＶＡＥ）を含むニューラルネットワークモデルを対象としている。
【００２０】
エンコーダ部分は、発話を表す言語的特徴を条件とする多数の基準オーディオ信号を符号化することによって、韻律を表す発話埋め込みをトレーニングすることができる。言語的特徴は、各音素の個々の音、各音節が強勢を有するかまたは無強勢であるか、発話における各単語のタイプ（例えば、名詞／形容詞／動詞）および／または単語の位置、ならびに発話が質問であるかまたはフレーズであるかを含み得るが、これらに限定されない。各発話埋め込みは、固定長の数値ベクトルによって表される。いくつかの実装形態では、固定長の数値ベクトルは、２５６に等しい値を含む。しかしながら、他の実装形態では、２５６より大きいまたは小さい値（例えば、１２８）を有する固定長の数値ベクトルが使用されてもよい。デコーダ部分は、固定長発話埋め込みを、第１のデコーダを介して音節持続時間のシーケンスに復号化するとともに、音節持続時間を使用してピッチおよびエネルギーの固定長フレーム（例えば、５ミリ秒）のシーケンスに復号化することができる。トレーニング中、デコーダ部分によって予測された音節持続時間と、ピッチおよびエネルギーの固定長フレームとは、固定長発話埋め込みに関連付けられた基準オーディオ信号からサンプリングされた音節持続時間と、ピッチおよびエネルギーの固定長フレームとに厳密に一致する。

【００２１】
本開示のＶＡＥは、長短期記憶（ＬＳＴＭ）ブロックの階層的なスタックされた層を組み込んだクロックワーク階層型変分オートエンコーダ（Ｃｌｏｃｋｗｏｒｋ Ｈｉｅｒａｒｃｈａｌ Ｖａｒｉａｔｉｏｎａｌ Ａｕｔｏｅｎｃｏｄｅｒ，ＣＨｉＶＥ）を含み、ＬＳＴＭブロックの各層は、発話の構造を組み込んでいる。ここで、各ＬＳＴＭブロックは、１つまたは複数のＬＳＴＭセルに分割することができる。発話は、音素、音節、単語、フレーズ、または文などの言語ユニットのいずれか１つまたは組み合わせに分割できるため、ＬＳＴＭブロックは、そのようなユニットを表す１つまたは複数の層を含み得る。例えば、ＬＳＴＭブロックは、音素、音節、単語、フレーズ、または文を表す１つまたは複数の層を含むＬＳＴＭセルを含む。さらに、ＬＳＴＭセルのスタックされた層の階層は、階層的入力データの長さに合わせて可変的にクロックされる。例えば、入力データが、３音節の単語と、それに続く４音節の単語を含む場合、ＣＨｉＶＥの音節層は、第１の入力単語について、単語層の単一クロックに対し３回クロックし、次いで、音節層は、第２の単語について、単語層の後続の単一クロックに対しさらに４回クロックする。このように、所与のＬＳＴＭセルに関連付けられたメモリが約０．５秒（すなわち、５ミリ秒のフレームレートで１００回ステップ）しか有効でなく、したがって、音声の２音節または３音節分のＬＳＴＭセルメモリしか提供できないフレームベースの技術を使うのではなく、ＣＨｉＶＥの音素、単語、および音節層は、それぞれ音素、単語、音節でクロックして、スタックされた層のＬＳＴＭセルに、過去１００単語、音節または音素にわたるメモリを与える。追加的または代替的に、ＣＨｉＶＥは、発話構造を表すために階層的な層にＬＳＴＭ構造を使用する代わりに、他の形式のニューラルネットワーク（ＮＮ）またはリカレントニューラルネットワーク（ＲＮＮ）を使用するように適合されていてもよい。

【００２２】
推論中、ＣＨｉＶＥは、テキスト発話を受信し、テキスト発話のために発話埋め込みを選択するように構成されている。受信されたテキスト発話は、少なくとも１つの単語を有し、各単語は、少なくとも１つの音節を有し、各音節は、少なくとも１つの音素を有している。テキスト発話は、発話から合成音声を生成するための適切な韻律を導くためのコンテキスト、セマンティック情報、および語用論情報を欠いているため、ＣＨｉＶＥは、その選択された発話埋め込みを潜在変数として使用して、意図された韻律を表す。その後、ＣＨｉＶＥは、選択された発話埋め込みを使用して、音節に含まれる各音素の言語的特徴を、その音節に対する対応する韻律音節埋め込みで符号化することにより、各音節の持続時間を予測し、音節に対する予測された持続時間に基づいて各音節のピッチを予測する。最後に、ＣＨｉＶＥは、各固定長ピッチフレームが音節の予測ピッチを表すように、各音節の予測された持続時間に基づいて複数の固定長ピッチフレームを生成するように構成されている。ＣＨｉＶＥは、同様に、音節の予測された持続時間に基づいて各音節のエネルギー（例えば、ラウドネス）を予測し、各々が音節の予測されたエネルギーを表す複数の固定長エネルギーフレームを生成し得る。固定長ピッチおよび／またはエネルギーフレームは、ＴＴＳシステムのユニット選択モデルまたはウェーブネットモデルに提供されて、入力固定長発話埋め込みによって提供される意図された韻律を備えた合成音声を生成することができる。

【００２３】
一般に、いくつかの音声合成システムは、２つのフェーズに分けられる場合がある。発話に含まれる音声に影響を与えるファクターを識別する言語仕様を生成する第１フェーズ、および言語仕様を使用して合成音声の波形を生成する第２フェーズである。いくつかの例では、音声合成システムの別の側面が音声を生成するために使用する言語特性を予測する代わりに、ＣＨｉＶＥは、スペクトルフレーム（例えば、メルフレームなどの複数の固定長スペクトルフレーム）を予測するように代替的または追加的に構成されている。スペクトルフレームを予測することにより、ＣＨｉＶＥは、言語仕様から波形を生成するための音声合成システムのさらなる処理を最小化することができる。ここで、ＣＨｉＶＥは、選択された発話埋め込みを使用して、音節に含まれる各音素の言語的特徴を、その音節に

対する対応する韻律音節埋め込みで符号化することにより、各音節の持続時間を予測する。各音節に対する予測された持続時間を使用して、ＣＨｉＶＥは、複数の固定長スペクトルフレームを生成することができる。例えば、第１のデコーダは、各音節に対する予測された持続時間を生成し、これは、第２のデコーダが各音節に対して生成すべきスペクトルフレームの数を示す。次に、第２のデコーダは、各音節に対する予測された持続時間によって示されるスペクトルフレームの数をシーケンス復号化する。次に、スペクトルフレームは、音声合成システムによって使用されて、合成音声を生成することができる。例えば、スペクトルフレームは、ニューラルボコーダに提供される。

【００２４】

残念ながら、特定の言語ユニットに依存すると、ピッチ、エネルギー、またはスペクトルフレームのような言語ファクターを予測するのに支障が生じる可能性がある。特に、音声サンプル（例えば、音声合成システムをトレーニングする録音されたサンプル）は、音素シーケンスが不正確な可能性がある。これらの不正確さのために、音節アライメント（例えば、音節の境界）を使用し、次いで音素の境界を使用する方が、信頼性が高い場合がある。音素の場合、音声サンプルから予想される各音素を明確に定義するのは容易ではなく単純でもない場合がある。人は３つの音素を発音しようとし得るが、文脈やその人の方言によっては、各音素を明確または正確に発音できない場合がある。収集されたサンプルでは、サンプルを提供する人によって、または同じ人からのサンプルの間で、音素（音素レベル）の方言に違いがある場合がある。同じ人でも、文章またはその人のコミュニケーション速度に応じて単語を異なって発音する場合がある。素早く話す場合、人は音素をブレンドしたり、または音素を完全に脱落させたりすることがある。

【００２５】

別の例として、しばしば、地域の方言では、その地域内でよく使用される単語の発音が明確でない場合がある。説明のために、これは都市名で起こり得る。例えば、エジンバラ（Ｅｄｉｎｂｕｒｇｈ）市には少なくとも３つの音節、すなわち「Ｅｄ－ｉｎ－ｂｕｒｇｈ」を有しているが、早口や地元の方言で表現すると、「Ｅｍ－ｂｒａ」と発音され、基本的には２音節に短縮される。これらの２つの音節では、エジンバラの発音に存在すると予想されるすべての音素を一致させることは困難である（例えば、３音節の形式）。これらの発音の違いは、短い基本的な単語または複数の音節を持つ大きな単語で発生する可能性がある。例えば、「ｔｈ－ａｈｈ－ｔ」の発音を有する「ｔｈａｔ」のような単純な単語は、「ｔｈｕｈｔ」に短縮される場合がある。音声の発音が最適ではない場合（例えば、曖昧になったり急いだりした場合）、サンプルに存在する音素シーケンスは必ずしも正確ではない。モデルがこれらのタイプの不正確な音素サンプルに音素を割り当てようとした場合、モデルはエラーを起こす危険性がある。

【００２６】

信頼できないデータの原因となるこれらの音素の問題のいくつかを克服するために、ＣＨｉＶＥは、音素アライメントに依存しないアプローチを使用することができる。例えば、ＣＨｉＶＥは、音素の数に依存せず、音節のフレーム数を予測しようとする。ここで、ＣＨｉＶＥは、音素ごとの個々のフレーム数について何も知らない可能性がある。いくつかの例では、音節の音声コンテンツに関する情報は、音素の言語的特徴（例えば、音節に関する言語情報）に基づいている。これらの例では、ＣＨｉＶＥは、音節情報（例えば、音節あたりのフレーム数）を予測するために、符号化されてアテンション機構に提供される音素の言語的特徴を使用する。

【００２７】

図１は、制御可能な韻律モデル３００を提供するためにディープニューラルネットワーク２００をトレーニングし、韻律モデル３００を使用してテキスト発話３１０の韻律表現３０２を予測するための例示的なシステム１００を示す。システム１００は、データ処理ハードウェア１１２、およびデータ処理ハードウェア１１２と通信し、データ処理ハードウェア１１２に動作を実行させる命令を格納しているメモリハードウェア１１４を有するコンピューティングシステム１１０を含む。いくつかの実装形態では、コンピューティン

グシステム１１０（例えば、データ処理ハードウェア１１２）は、入力テキスト発話３１０からの合成音声１２２の韻律を制御するためにテキスト音声変換（ｔｅｘｔ－ｔｏ－ｓｐｅｅｃｈ，ＴＴＳ）システム１２０に、トレーニングされたディープニューラルネットワーク２００に基づく韻律モデル３００を提供する。入力テキスト発話３１０は、合成音声１２２の適切な韻律を導くためのコンテキスト、セマンティクス、および語用論を伝達する方法を有しないので、韻律モデル３００は、テキスト発話３１０から抽出された言語的特徴でモデル３００を条件付け、固定長発話埋め込み２０４を、テキスト発話３１０の意図する韻律を表す潜在変数として使用することによって、入力テキスト発話３１０の韻律表現３０２を予測することができる。いくつかの例では、コンピューティングシステム１１０は、ＴＴＳシステム１２０を実装する。他の例では、コンピューティングシステム１１０およびＴＴＳシステム１２０は、別個であり、互いに物理的に分離されている。コンピューティングシステム１２０は、分散システム（例えば、クラウドコンピューティング環境）を含み得る。

【００２８】

いくつかの実装形態では、ディープニューラルネットワーク２００は、基準オーディオ信号２０２の大きなセットでトレーニングされる。各基準オーディオ信号２０２は、マイクロフォンによって録音された、韻律表現を有する人間音声の音声発話を含み得る。トレーニング中、ディープニューラルネットワーク２００は、同じ音声発話に対して、異なる韻律を有する複数の基準オーディオ信号２０２を受信することができる（すなわち、同じ発話を複数の異なる方法で話すことができる）。ここで、基準オーディオ信号２０２は、コンテンツが同じであっても、音声発話の持続時間が変化するように可変長である。ディープニューラルネットワーク２００は、各基準オーディオ信号２０２に関連付けられた韻律表現を、対応する固定長の発話埋め込み２０４に符号化／圧縮するように構成されている。ディープニューラルネットワーク２００は、各固定長発話埋め込み２０４を、発話埋め込み２０４に関連付けられた基準オーディオ信号２０２の対応するトランスクリプト２０６とともに、（例えば、コンピューティングシステム１１０のメモリハードウェア１１４上の）発話埋め込みストレージ１３０に格納することができる。ディープニューラルネットワーク２００は、トランスクリプト２０６から抽出された言語的特徴を条件とする固定長発話埋め込み２０４を逆伝播して、各音節のピッチ、エネルギー、および持続時間の固定長フレームを生成することによってさらにトレーニングされ得る。

【００２９】

推論中、コンピューティングシステム１１０は、韻律モデル３００を使用して、テキスト発話３１０の韻律表現３０２を予測することができる。韻律モデル３００は、テキスト発話３１０のための発話埋め込み２０４を選択することができる。発話埋め込み２０４は、テキスト発話３１０の意図された韻律を表す。以下、図２Ａ～図２Ｃおよび図３Ａ～図３Ｄを参照してより詳細に説明されるように、韻律モデル３００は、選択された発話埋め込み２０４を使用して、テキスト発話３１０の韻律表現３０２を予測することができる。韻律表現３０２は、テキスト発話３１０の予測されたピッチ、予測されたタイミング、および予測されたラウドネス（例えば、エネルギー）を含み得る。示されている例では、ＴＴＳシステム１２０は、韻律表現３０２を使用して、テキスト発話３１０から意図された韻律を有する合成された音声１２２を生成する。

【００３０】

図２Ａ～図２Ｃは、韻律の制御可能なモデルを提供するクロックワーク階層型変分オートエンコーダ（ＣＨｉＶＥ）３００（「オートエンコーダ３００」）の階層的言語構造（例えば、図１のディープニューラルネットワーク）２００を示す。韻律の制御可能なモデルは、所与の入力テキストの各音節について、所与の入力テキストからの固有のマッピングまたは他の言語仕様に依存することなく、音節の持続時間、音節のピッチ（Ｆ０）およびエネルギー（Ｃ０）輪郭を共同で予測して、意図された／選択された韻律を有する合成音声１２２を生成することができる。オートエンコーダ３００は、基準オーディオ信号２０２からサンプリングされた複数の固定長基準フレーム２１０を固定長発話埋め込み２０

４に符号化するエンコーダ部分３２０（図２Ａ）と、固定長発話埋め込み２０４を復号化する方法を学習するデコーダ部分３３０（図２Ｂおよび図２Ｃ）とを含む。デコーダ部分３３０は、固定長発話埋め込み２０４を複数の固定長予測フレーム（ｆｉｘｅｄ－ｌｅｎｇｔｈ　ｐｒｅｄｉｃｔｅｄ　ｆｒａｍｅ）２６０に復号化することができる（例えば、発話埋め込み２０４に対してピッチ（Ｆ０）、エネルギー（Ｃ０）、またはスペクトル特性を予測するため）。これから明らかになるように、オートエンコーダ３００は、デコーダ部分３３０から出力される予測フレーム２６０の数が、エンコーダ部分３２０に入力される基準フレーム２１０の数と等しくなるようにトレーニングされる。さらに、オートエンコーダ３００は、基準フレーム２１０および予測フレーム２６０に関連付けられたデータが互いに実質的に一致するようにトレーニングされる。

【００３１】

図２Ａを参照すると、エンコーダ部分３２０は、入力基準オーディオ信号２０２から固定長基準フレーム２１０のシーケンスを受信する。入力基準オーディオ信号２０２は、ターゲット韻律を含む、マイクロフォンによって記録された人間音声の音声発話を含み得る。エンコーダ部分３２０は、同じ音声発話に対して、異なる韻律を有する複数の基準オーディオ信号２０２を受信することができる（すなわち、同じ発話を複数の異なる方法で話すことができる）。例えば、同じ音声発話でも、話された基準が質問への回答である場合は、音声発話が質問である場合と比較して、韻律が異なり得る。基準フレーム２１０は、各々、５ミリ秒（ｍｓ）の持続時間を含み、基準オーディオ信号２０２に対するピッチの輪郭（Ｆ０）またはエネルギーの輪郭（Ｃ０）のうちの１つを表すことができる。並行して、エンコーダ部分３２０は、各々が５ミリ秒の持続時間を含み、基準オーディオ信号２０２に対するピッチの輪郭（Ｆ０）またはエネルギーの輪郭（Ｃ０）のうちの他の１つを表す基準フレーム２１０の第２のシーケンスを受信してもよい。したがって、基準オーディオ信号２０２からサンプリングされたシーケンス基準フレーム２１０は、持続時間、ピッチ輪郭、および／またはエネルギー輪郭を提供して、基準オーディオ信号２０２に対する韻律を表す。基準オーディオ信号２０２の長さまたは持続時間は、基準フレーム２１０の総数の合計と相関している。

【００３２】

エンコーダ部分３２０は、互いに対して相対的にクロックする、基準オーディオ信号２０２に対する基準フレーム２１０、音素２２０，２２０ａ、音節２３０，２３０ａ、単語２４０，２４０ａ、および文２５０，２５０ａの階層レベルを含む。例えば、基準フレーム２１０のシーケンスに関連付けられたレベルは、音素２２０のシーケンスに関連付けられた次のレベルよりも速くクロックする。同様に、音節のシーケンス２３０に関連付けられたレベルは、音素３３０のシーケンスに関連付けられたレベルよりも遅く、かつ単語のシーケンス２４０に関連付けられたレベルよりも速くクロックする。したがって、より遅くクロックする層は、入力として、より速くクロックする層からの出力を受け取り、その結果、より速い層の最終クロック（すなわち、状態）の後の出力は、対応するより遅い層への入力として取られ、本質的にシーケンス間（ｓｅｑｕｅｎｃｅ－ｔｏ－ｓｅｑｕｅｎｃｅ）エンコーダを提供する。示されている例では、階層レベルは、長短期記憶（ＬＳＴＭ）レベルを含む。

【００３３】

図２Ａは、基準オーディオ信号２０２に対する階層レベルの例を示している。この例では、基準オーディオ信号２０２は、３つの単語２４０，２４０Ａ～Ｃを有する１つの文２４０，２４０Ａを含む。第１の単語２４０，２４０Ａは、２つの音節２３０，２３０Ａａ～Ａｂを含む。第２の単語２４０，２４０Ｂは、１つの音節２３０，２３０Ｂａを含む。第３の単語２４０，２４０ａは、２つの音節２３０，２３０Ｃａ～Ｃｂを含む。第１の単語２４０，２４０Ａの第１の音節２３０，２３０Ａａは、２つの音素２２０，２２０Ａａ１～Ａａ２を含む。第１の単語２４０，２４０Ａの第２の音節２３０，２３０Ａｂは、１つの音素２２０，２２０Ａｂ１を含む。第２の単語２４０，２４０Ｂの第１の音節２３０，２３０Ｂａは、３つの音素２２０，２２０Ｂａ１～Ｂａ３を含む。第３の単語２４０，

２４０Ｃの第１の音節２３０，２３０Ｃａは、１つの音素２２０，２２０Ｃａ１を含む。第３の単語２４０，２４０Ｃの第２の音節２３０，２３０Ｃｂは、２つの音素２２０，２２０Ｃｂ１～Ｃｂ２を含む。

【００３４】

いくつかの例では、エンコーダ部分３２０は、最初に、基準フレーム２１０のシーケンスをフレームベースの音節埋め込み２３２，２３２Ａａ～Ｃｂに符号化する。いくつかの実装形態では、基準フレーム２１０は、音素２２０Ａａ１～２２０Ｃｂ２のシーケンスを定義する。ここで、基準フレーム２１０のサブセットを１つまたは複数の音素２２０に符号化する代わりに、エンコーダ部分３２０は、代わりに、単音（ｐｈｏｎｅ）レベルの言語的特徴２２２，２２２Ａａ１～Ｃｂ２を単音特徴ベースの音節埋め込み２３４，２３４Ａａ～Ｃｂに符号化することによって音素２２０を説明する。それぞれの音節埋め込み２３２，２３４は、対応する音節２３０に関連付けられた持続時間、ピッチ（Ｆ０）、および／またはエネルギー（Ｃ０）を示す数値ベクトルを参照することができる。さらに、各音節埋め込み２３２，２３４は、音節２３０のレベルに対して対応する状態を示している。

【００３５】

図２Ａを参照すると、階層的な層の斜めのハッチングパターンを含むブロックは、階層の特定のレベルの言語的特徴に対応する。フレームベースの音節埋め込み２３２および単音特徴ベースの音節埋め込み２３４を用いて、エンコーダ部分３２０は、これらの音節埋め込み２３２，２３４を他の言語的特徴とともに符号化する。例えば、エンコーダ部分３２０は、音節レベルの言語的特徴２３６，２３６Ａａ～Ｃｂ、単語レベルの言語的特徴２４２，２４２Ａ～Ｃ、および／または文レベルの言語的特徴２５２，２５２Ａを用いて、音節埋め込み２３２，２３４を符号化する。言語的特徴２３６，２４２，２５２を用いて音節埋め込み２３２，２３４を符号化することにより、エンコーダ部分３２０は、基準オーディオ信号２０２のための発話埋め込み２０４を生成する。発話埋め込み２０４は、基準オーディオ信号２０４のそれぞれのトランスクリプト２０６（例えば、テキスト表現）と共にデータストレージ１３０（図１）に格納され得る。トランスクリプト２０６から、言語的特徴２２２，２３６，２４２，２５２を抽出して、階層的言語構造２００のトレーニングを調整する際に使用するために格納することができる。言語的特徴（例えば、言語的特徴２２２，２３６，２４２，２５２）は、各音素の個々の音、各音節が強勢を有するかまたは無強勢であるか、発話における各単語のタイプ（例えば、名詞／形容詞／動詞）および／または単語の位置、ならびに発話が質問であるかまたはフレーズであるかを含み得るが、これらに限定されない。

【００３６】

図２Ａの例では、符号化ブロック３２２，３２２Ａａ～Ｃｂは、言語的特徴２３６，２４２，２５２と音節埋め込み２３２，２３４との組み合わせを描写するために示されている。ここで、ブロック３２２は、音節のレートでシーケンス符号化されて、発話埋め込み２０４を生成する。例として、第１のブロック３２２Ａａは、第２のブロック３２２Ａｂへの入力として供給される。第２のブロック３２２Ａｂは、第３のブロック３２２Ｂａへの入力として供給される。第３のブロック３２２Ｃａは、第４のブロック３２２Ｃａへの入力として供給される。第４のブロック３２２Ｃａは、第５のブロック３２２Ｃｂに供給される。いくつかの構成では、発話埋め込み２０４は、各基準オーディオ信号２０２の平均μおよび標準偏差σを含み、ここで、平均μおよび標準偏差σは、複数の基準オーディオ信号２０２のトレーニングデータに関するものである。

【００３７】

いくつかの実装形態では、各音節２３０は、入力として、基準フレーム２１０のサブセットの対応するエンコーディングを受け取り、符号化されたサブセットにおける基準フレーム２１０の数に等しい持続時間を含む。示されている例では、最初の７つの固定長基準フレーム２１０は、音節２３０Ａａに符号化され、次の４つの固定長基準フレーム２１０は、音節２３０Ａｂに符号化され、次の１１個の固定長基準フレーム２１０は、音節２３０Ｂａに符号化され、次の３つの固定長基準フレーム２１０は、音節２３０Ｃａに符号化

され、最後の６つの固定長基準フレーム２１０は、音節２３０Ｃｂに符号化される。したがって、音節２３０のシーケンスの各音節２３０は、音節２３０に符号化された基準フレーム２１０の数および対応するピッチおよび／またはエネルギー輪郭に基づく対応する持続時間を含むことができる。例えば、音節２３０Ａａは、３５ミリ秒に等しい持続時間を含み（すなわち、各々が５ミリ秒の固定長を有する６つの基準フレーム２１０）、音節２３０Ａｂは、２０ミリ秒に等しい持続時間を含む（すなわち、各々が５ミリ秒の固定長を有する４つの基準フレーム２１０）。したがって、基準フレーム２１０のレベルは、音節２３０のレベルにおける音節２３０Ａａと次の音節２３０Ａｂとの間の単一のクロッキングに対して合計１０回クロックする。音節２３０の持続時間は、音節２３０のタイミング、および隣接する音節２３０の間の休止を示すことができる。

【００３８】

いくつかの実装形態では、エンコーダ部分３２０によって生成された発話埋め込み２０４は、基準オーディオ信号２０２の韻律を表す数値ベクトルを含む固定長の発話埋め込み２０４である。いくつかの例では、固定長発話埋め込み２０４は、「１２８」または「２５６」に等しい値を有する数値ベクトルを含む。エンコーダ部分３２０は、各々が同じ音声発話／フレーズに対応するが、異なる韻律を有する複数の基準オーディオ信号２０２を符号化することができ、すなわち、各基準オーディオ信号２０２は、同じ発話を伝達するが、異なって話される。

【００３９】

図２Ｂおよび図２Ｃを参照すると、変分オートエンコーダ３００のデコーダ部分３３０は、発話の韻律を表す発話埋め込み２０４を最初に復号化することによって、複数の音節埋め込み２３２，２３４（例えば、固定長音節埋め込み）を生成するように構成されている。トレーニング中、発話埋め込み２０４は、基準オーディオ信号２０２からサンプリングされた複数の固定長基準フレーム２１０を符号化することによって、図２Ａのエンコーダ部分３２０から出力された発話埋め込み２０４を含み得る。したがって、デコーダ部分３３０は、トレーニング中に発話埋め込み２０４を逆伝播して、複数の固定長基準フレーム２１０に厳密に一致する複数の固定長予測フレーム２６０を生成するように構成されている。例えば、ピッチ（Ｆ０）およびエネルギー（Ｃ０）の両方に対する固定長予測フレーム２６０は、トレーニングデータとしてエンコーダ部分３２０に入力される基準オーディオ信号２０２の基準韻律と実質的に一致するターゲット韻律（例えば、予測された韻律）を表すために並行して生成されてもよい。追加的または代替的に、固定長予測フレーム２６０は、ＴＴＳシステム１２０（図１）に提供され得るスペクトルフレーム（例えば、メルフレーム）であってよい。いくつかの例では、ＴＴＳシステム１２０（図１）は、固定長予測フレーム２６０を使用して、固定長発話埋め込み２０４に基づいて、選択された韻律を有する合成音声１２２を生成する。例えば、ＴＴＳシステム１２０のユニット選択モジュール、ＷａｖｅＮｅｔモジュール、またはニューラルボコーダは、フレーム２６０を使用して、意図された韻律を有する合成音声１３２を生成することができる。

【００４０】

示されている例では、デコーダ部分３３０は、エンコーダ部分３２０（図２Ａ）から受信された発話埋め込み２０４（例えば、「２５６」または「１２８」の数値）を階層レベルに復号化する。例えば、階層レベルは、文２５０，２５０ｂ、単語２４０，２４０ｂ、音節２３０，２３０ｂ、音素２２０，２２０ｂ、および固定長予測フレーム２６０に対応するレベルを含む。具体的には、固定長発話埋め込み２０４は、デコーダ部分３３０に対する階層入力データの変分層に対応し、スタックされた階層レベルの各々は、階層入力データの長さに可変的にクロックされる長短期記憶（ＬＳＴＭ）処理ブロックを含む。例えば、音節レベル２３０は、単語レベル２４０よりも速く、かつ音素レベル２２０よりも遅くクロックする。各レベルの長方形のブロックは、それぞれの文、単語、音節、音素、またはフレームの１つまたは複数のＬＳＴＭ処理セルに対応する。有利には、オートエンコーダ３００は、単語レベル２４０のＬＳＴＭ処理セルに最後の１００単語にわたるメモリを与え、音節レベル２３０のＬＳＴＭセルに最後の１００音節にわたるメモリを与え、音

素レベル２２０のＬＳＴＭセルに最後の１００音素にわたるメモリを与え、固定長ピッチおよび／またはエネルギーフレーム２６０のＬＳＴＭセルに最後の１００個の固定長フレーム２６０にわたるメモリを与える。固定長フレーム２６０が、それぞれ５ミリ秒の持続時間（例えば、フレームレート）を含む場合、対応するＬＳＴＭ処理セルは、最後の５００ミリ秒（例えば、０．５秒）にわたってメモリを提供する。

【００４１】

図２Ｂおよび図２Ｃを参照すると、いくつかの例では、階層的言語構造２００のデコーダ部分３３０は、エンコーダ部分３２０によって符号化された固定長発話埋め込み２０４を逆伝播する。例えば、図２Ｂは、階層的言語構造２００のデコーダ部分３３０が、固定長発話埋め込み２０４を、１つの文２５０，２５０Ａ、３つの単語２４０Ａ～２４０Ｃのシーケンス、５つの音節２３０Ａａ～２３０Ｃｂのシーケンス、および９つの音素２３０Ａａ１～２３０Ｃｂ２のシーケンスに逆伝播して、予測固定長フレーム２６０のシーケンスを生成することを示している。図２Ｃのようないくつかの実装形態では、デコーダ部分３３０は、音節レベル２３０から音素レベル２２０に逆伝播するのではなく、音節レベル２３０からフレームレベル２１０に逆伝播する。ここで、このアプローチは、音素アラインメントがモデル２００にとって問題となる可能性がある状況において、予測されるフレーム２６０の精度を高めることができる。デコーダ部分３３０は、入力テキストの言語的特徴（例えば、言語的特徴２２２，２３６，２４２，２５２）に基づいて条件付けられる。より速くクロックする層からの出力が、より遅くクロックする層によって入力として受け取られ得るエンコーダ部分３２０（例えば、図２Ａに示されるような）とは対照的に、デコーダ部分３３０は、より速くクロックする層に供給する、より遅くクロックする層からの出力を含み、これにより、より遅くクロックする層の出力が、それに付加されたタイミング信号を用いて、各クロックサイクルで、より速くクロックする層の入力に分配される。

【００４２】

図３Ａ～図３Ｃを参照すると、いくつかの実装形態では、オートエンコーダ３００は、階層的言語構造２００を使用して、推論中に所与のテキスト発話３１０の韻律表現を予測する。例えば、オートエンコーダ３００は、所与のテキスト発話３１０の各音節２３０に対して、音節２３０の持続時間およびピッチＦ０および／またはエネルギーＣ０輪郭を一緒に予測することによって、所与のテキスト発話３１０に対する韻律表現を予測する。テキスト発話３１０は、テキスト発話３１０の適切な韻律を示すためのコンテキスト、セマンティック情報、または語用論情報を提供しないので、オートエンコーダ３００は、潜在変数として発話埋め込み２０４を選択して、テキスト発話３１０の意図された韻律を表す。

【００４３】

発話埋め込み２０４は、発話埋め込みデータストレージ１３０（図１）から選択することができる。ストレージ１３０内の各発話埋め込み２０４は、トレーニング中に、対応する可変長基準オーディオ信号２０２（図２Ａ）からエンコーダ部分３２０（図２Ａ）によって符号化され得る。具体的には、エンコーダ部分３１０は、トレーニング中に可変長基準オーディオ信号２０２の韻律を、固定長発話埋め込み２０４に圧縮し、推論時にデコーダ部分３３０によって使用するため、各発話埋め込み２０４を、対応する基準オーディオ信号２０２のトランスクリプト２０６と共に、発話埋め込みデータストレージ１３０に格納する。示される例では、オートエンコーダ３００は、最初に、テキスト発話３１０に厳密に一致するトランスクリプト２０６を有する発話埋め込み２０４を見つけ、次に、発話埋め込み２０４のうちの１つを選択して、所与のテキスト発話３１０の韻律表現３０２（図１）を予測することができる。いくつかの例では、固定長の発話埋め込み２０４は、ターゲット韻律の特定のセマンティクスおよび語用論を表す可能性が高い埋め込み２０４の潜在空間内の特定の点を選ぶことによって選択される。他の例では、潜在空間は、テキスト発話３１０の意図された韻律を表すために、ランダムな発話埋め込み２０４を選択するためにサンプリングされる。さらに別の例では、オートエンコーダ３００は、テキスト発話３１０の言語的特徴に対する最も可能性の高い韻律を表すために、厳密に一致するトラ

ンスクリプト２０６を有する発話埋め込み２０４の平均を選択することによって、潜在空間を多次元ユニットガウシアンとしてモデル化する。例えば、オートエンコーダ３００は、発話埋め込み２０４を選択してテキスト発話３１０に対する韻律表現３０２を生成する場合に、各発話埋め込み２０４に関連付けられた平均μおよび／または標準偏差σを使用する。トレーニングデータの韻律の変化が適度に中立である場合、発話埋め込み２０４の平均を選択する最後の例は妥当な選択である。

【００４４】

図３Ａ～図３Ｃは、階層的言語構造２００の単語レベル２４０で表される３つの単語２４０Ａ，２４０Ｂ，２４０Ｃを有するテキスト発話３１０を示している。第１の単語２４０Ａは、音節２３０Ａａ，２３０Ａｂを含む。第２の単語２４０Ｂは、１音節２３０Ｂａを含む。第３の単語２４０Ｃは、音節２３０Ｃａ，２３０Ｃｂを含む。したがって、階層的言語構造２００の音節レベル２３０は、テキスト発話３１０の５つの音節２３０Ａａ～２３０Ｃｂのシーケンスを含む。ＬＳＴＭ処理セルの音節レベル２３０において、オートエンコーダ３００は、第１の音節埋め込み（例えば、フレームベースの音節埋め込み２３２）および第２の音節埋め込み（例えば、単音特徴ベースの音節埋め込み２３４）を生成／出力するように構成されている。第１の音節埋め込み２３２Ａａ，２３２Ａｂ，２３２Ｂａ，２３２Ｃａ，および２３２Ｃｂを出力するために、オートエンコーダ３００は、以下の入力、すなわち、固定長発話埋込み２０４、テキスト発話３１０に関連付けられた発話レベルの言語的特徴（例えば、文レベルの言語的特徴２５２）、音節２３０を含む単語２４０に関連付けられた単語レベルの言語的特徴２４２、音節２３０に対する音節レベルの言語的特徴２３６を使用する。第２の音節埋め込み２３４Ａａ，２３４Ａｂ，２３４Ｂａ，２３４Ｃａ，および２３４Ｃｂを出力するために、オートエンコーダ３００は、音節２３０に関連付けられた単音レベルの言語的特徴２２２を使用する。発話レベルの言語的特徴２５２は、テキスト発話３２０が質問であるかどうか、質問への回答であるかどうか、フレーズであるかどうか、文であるかどうかなどを含むがこれらに限定されない。いくつかの例では、Ｄｖｅｃｔｏｒまたは他の複雑な話者アイデンティティ表現が、発話レベルの言語的特徴２５２として含まれ得る。単語レベルの言語的特徴２４２は、これらに限定されないが、単語タイプ（例えば、名詞、代名詞、動詞、形容詞、副詞など）およびテキスト発話３１０における単語の位置を含み得る。音節レベルの言語的特徴２３６は、これに限定されないが、音節２３０が強勢を有するか無強勢であるかを含み得る。いくつかの実装形態では、Ｄｖｅｃｔｏｒまたは他の複雑な話者アイデンティティ表現が、発話レベル（例えば、単語レベル２４０または音節レベル２３０）より下のレベルにおける言語的特徴として含まれ得る。

【００４５】

示されている例では、音節レベル２３０内の各音節２３０Ａａ，２３０Ａｂ，２３０Ｂａ，２３０Ｃａ，２３０Ｃｂは、個々の固定長予測ピッチ（Ｆ０）フレーム２６０，２６０Ｆ０（図３Ａ）を復号化するため、個々の固定長予測エネルギー（Ｃ０）フレーム２６０，２６０Ｃ０（図３Ｂ）を復号化するため、および／または個々の固定長スペクトル（Ｍ０）フレーム２６０，２６０Ｍ０を復号化するため、対応する音節埋め込み２３２Ａａ～Ｃｂ，２３４Ａａ～Ｃｂを出力する対応するＬＳＴＭ処理セルに関連付けられ得る。いくつかの実装形態では、オートエンコーダ３００は、２つ以上のタイプのフレームＦ０，Ｃ０，Ｍ０を並列に復号化する。図３Ａは、音節２３０に対する持続時間（タイミングおよび休止）およびピッチ輪郭を示す複数の固定長予測ピッチ（Ｆ０）フレーム２６０Ｆ０を含む、音節レベル２３０における各音節２３０を示している。ここで、持続時間およびピッチ輪郭は、音節２３０の韻律表現に対応する。図３Ｂは、音節２３０に対する持続時間およびエネルギー輪郭を示す複数の固定長予測エネルギー（Ｃ０）フレーム２６０Ｃ０を含む、音節レベル２４０における各音節２３０を示している。

【００４６】

音節レベル２３０の第１の音節２３０Ａａ（すなわち、ＬＳＴＭ処理セルＡａ）は、対応する音節埋め込み２３２Ａａを生成するための入力として、固定長発話埋め込み２０４

、テキスト発話３１０に関連付けられた発話レベルの言語的特徴２５２、第１の単語２３０Ａに関連付けられた単語レベルの言語的特徴２４２Ａ、および音節２３０Ａａに対する音節レベルの言語的特徴２３６Ａａを受け取る。音節レベル２３０の第２の音節２３０Ａｂは、対応する音節埋め込み２３２Ａｂを生成するための入力として、固定長発話埋め込み２０４、テキスト発話３１０に関連付けられた発話レベルの言語的特徴２５２、第１の単語２４０Ａに関連付けられた単語レベルの言語的特徴２４２Ａ、および音節２３０Ａｂに対する対応する音節レベルの言語的特徴２３６を受け取る。ここで、オートエンコーダ３００がテキスト発話３１０を音節埋め込み２３２，２３４に復号化する場合、これらの埋め込み２３２，２３４は、各音節２３０に対して状態３３２（例えば、状態３３２，３３２Ａａ～Ｃｂとして示される）を形成する。オートエンコーダ３００が後続の状態３３２を形成する場合、オートエンコーダ３００は、先行する音節２３０の状態も受け取る。換言すると、音節レベル２３０の各ＬＳＴＭ処理セルは、音節レベル２３０の直前のＬＳＴＭ処理セルの状態２３２を受け取る。いくつかの構成では、音節レベル２３０の各ＬＳＴＭ処理セルは、現在の音節レベル２３０のＬＳＴＭ処理セルに先行する各状態２３２を受け取る。例えば、第２の音節２３０Ａｂに関連付けられたＬＳＴＭ処理セルは、先行する第１の音節２３０Ａａの状態３３２，３３２Ａａを受け取る。図３Ａ～図３Ｃに示されるように、オートエンコーダ３００は、同様の方法で、音節レベル２３０における音節２３０Ｂａ，２３０Ｃａ，２３０Ｃｂの残りのシーケンスに対して、対応する音節埋め込み２３２Ｂａ～Ｃｂ，２３４Ｂａ～ＣＢを生成する。追加的または代替的に、オートエンコーダ３００は、連結によって、１つまたは複数の音節埋め込み２３２，２３４を用いて、より高いレベルの言語的特徴（例えば、文の特徴２５２、単語の特徴２４２など）を復号化することができる。

【００４７】

図３Ｄを参照すると、階層的言語構造２００の音素レベル２２０は、９つの音素２２０Ａａ１～２２０Ｃｂ２のシーケンスを含む。いくつかの実装形態では、オートエンコーダ３００は、各音素２２０Ａａ１～２２０Ｃｂ２に関連付けられた音素レベルの言語的特徴２２２を、音素符号化状態２２４，２２４Ａａ１～Ｃａ２に符号化する。音素レベルの言語的特徴２２２は、対応する音素２２０の音のアイデンティティを含み得るがこれに限定されない。図３Ｄに示されるように、アテンション機構３４０は、符号化された音素状態２２４を形成する単音レベルの特徴２２２，２２Ａａ１～Ｃｂ２にアテンションを与えるように構成されている。このアプローチでは、オートエンコーダ３００は、音素持続時間に依存する必要もなく、音素持続時間を予測もする必要もなく、むしろ、単音レベルの言語的特徴２２２へのアテンションを有するアテンション機構３４０を使用して、各音節２３０の音節持続時間２３８を予測する。

【００４８】

一般的に言えば、アテンション機構３４０は、入力を出力と（例えば、スコアリングすることによって）相関させるアライメントモデルである。符号化された隠された状態では、アテンション機構３４０は、各出力（例えば、予測されたフレーム２６０）について、各隠された状態（例えば、符号化された音素状態２２４）がどれだけのアテンションが考慮されるべきかを定義する重みのセットを形成し得る。異なるアラインメントスコア関数を使用する異なるタイプのアテンション機構３４０があってもよい。これらのアテンション機構３４０のいくつかの例は、コンテンツベースのアテンション、加法（ａｄｄｉｔｉｖｅ）アテンション、場所ベースのアテンション、一般的（ｇｅｎｅｒａｌ）アテンション、ドット積（ｄｏｔ－ｐｒｏｄｕｃｔ）アテンション、およびスケーリングされたドット積アテンションを含む。アテンション機構３４０のより広いカテゴリーは、セルフアテンション、グローバル／ソフトアテンション、および／またはローカル／ハードアテンションを含む。いくつかの例では、アテンション機構３４０のアライメントスコアは、単一の隠れ層（例えば、符号化された音素状態２２４の音素層２２０）を備えたフィードフォワードネットワークによってパラメータ化される。これらの例では、フィードフォワードネットワークは、オートエンコーダ３００の他の部分と共同でトレーニングされ得る。

【００４９】
オートエンコーダ３００のアテンション機構３４０は、これらのタイプのアテンションモデルの任意のものに基づくことができる。換言すると、オートエンコーダ３００は、異なるアテンションアプローチに従って、単音レベルの言語的特徴２２２に基づいて隠された状態２２４をスコアリングするように設計され得る。いくつかの例では、アテンション機構３４０は、単調にシフトする、位置に敏感なアテンション（すなわち、位置ベースのアテンション）の形をとる。ここで、アテンション機構３００の位置に敏感なアテンションは、ウィンドウによって制約され得る。いくつかの構成では、ウィンドウは、オートエンコーダ３００によって処理されている現在の音節２３０の音素情報に対応する。他の構成では、ウィンドウは、オートエンコーダ３００によって処理されている現在の音節２３０に隣接する音節２３０（例えば、前または後）を含むように拡張され得る。いくつかの実装形態では、アテンション機構３４０は、セルフアテンション機構に基づくトランスフォーマー（ｔｒａｎｓｆｏｒｍｅｒ）である。アテンション機構３４０としてトランスフォーマーを使用する場合、アテンション機構３４０は、符号化された音素状態２２４における符号化された単音レベルの言語的特徴２２２ではなく、単音レベルの言語的特徴２２２を入力として使用し得る。

【００５０】
図３Ａ～図３Ｃを引き続き参照すると、いくつかの例では、オートエンコーダ３００は、２つのデコーダを使用する（例えば、円で表されている）。オートエンコーダ３００のこれらのデコーダの各々は、入力として音節埋め込み２３２，２３４を受け取ることができる。第１のデコーダ（例えば、第１の音節状態３３２）は、アテンション機構３４０のアテンションとともに、音節埋め込み２３２，２３４を使用して、特定の音節２３０に対するフレーム数を予測する。次に、その特定の音節２３０について、オートエンコーダ３００は、（例えば、第２の音節状態３３４の）第２のデコーダを使用して、（例えば、第１の音節状態３３２の）第１のデコーダによって予測されたフレーム数をシーケンス復号化する。予測フレーム２６０の数に基づいて、オートエンコーダ３００は、予測フレーム２６０を生成する。換言すると、各音節２３０に対してアテンション機構３４０を使用して、予測音節持続時間は、固定長予測ピッチ（Ｆ０）フレーム２６０Ｆ０（例えば、図３Ａに示されるような）の数、固定長予測エネルギー（Ｃ０）フレーム２６０Ｃ０（例えば、図３Ｂに示されるような）の数、および／または固定長予測スペクトル（Ｍ０）フレーム２６０Ｍ０（例えば、図３Ｃに示されるような）の数を決定し得る。図３Ａ～図３Ｃは単一のアテンション機構３４０を示しているが、これは、単に説明を簡略化するためである。テキスト発話３１０の各音節２３０について、オートエンコーダ３００は、状態３３２および状態３３４でアテンション機構３４０を使用して、音節２３０に対応する予測フレーム２６０を生成する。例えば、図３Ａ～図３Ｃは、点線のボックスと相互作用するアテンション機構３４０を示しており、オートエンコーダ３００が、各音節２３０の各状態３３２，３３４でアテンション機構３４０を使用することを示している。

【００５１】
アテンション機構３４０および音節埋め込み２３２Ａａ，２３４Ａａに基づいて、オートエンコーダ３００は、７つの予測フレーム２６０に対応する第１の音節２３０Ａａの音節持続時間２３８，２３８Ａａを予測する。第１の音節２３０Ａａに対する状態３３４，３３４Ａａで、オートエンコーダ３００は、７つの予測フレーム２６０を復号化する。第２の音節２３０Ａｂについて、アテンション機構３４０および音節埋め込み２３２Ａａ～Ａｂ，２３４Ａａ～Ａｂに基づいて、オートエンコーダ３００は、４つの予測フレーム２６０の音節持続時間２３８，２３８Ａｂを予測する。第２の音節２３０Ａｂに対する状態３３４，３３４Ａｂで、オートエンコーダ３００は、４つの予測フレーム２６０を復号化する。第３の音節２３０Ｂａについて、アテンション機構３４０および音節埋め込み２３２Ａａ～Ｂａ，２３４Ａａ～Ｂａに基づいて、オートエンコーダ３００は、１１個の予測フレーム２６０の音節持続時間２３８，２３８Ｂａを予測する。第３の音節２３０Ｂａに対する状態３３４，３３４Ｂａで、オートエンコーダ３００は、１１個の予測フレーム２

６０を復号化する。第４の音節２３０Ｃａについて、アテンション機構３４０および音節埋め込み２３２Ａａ～Ｃａ，２３４Ａａ～Ｃａに基づいて、オートエンコーダ３００は、３つの予測フレーム２６０の音節持続時間２３８，２３８Ｃａを予測する。第４の音節２３０Ｃａに対する状態３３４，３３４Ｃａで、オートエンコーダ３００は、３つの予測フレーム２６０を復号化する。第５の音節２３０Ｃｂについて、アテンション機構３４０および音節埋め込み２３２Ａａ～Ｃｂ，２３４Ａａ～Ｃｂに基づいて、オートエンコーダ３００は、６つの予測フレーム２６０の音節持続時間２３８，２３８Ｃｂを予測する。第５の音節２３０Ｃｂに対する状態３３４，３３４Ｃｂで、オートエンコーダ３００は、６つの予測フレーム２６０を復号化する。

【００５２】

ここで、このシステム１００は、（例えば、音素レベルが階層的言語構造２００に直接影響を与えることなく）フレーム２６０を予測するために、オートエンコーダ３００を使用して音節２３０に焦点を合わせる。しかしながら、本明細書のアプローチは、階層的言語構造２００の他の層（例えば、文２５０、単語２４０など）に焦点を合わせるように適合されてもよい。例えば、オートエンコーダ３００は、フレーム２６０を予測するために単語２４０または文２５０に焦点を合わせる。これらの他のアプローチでは、１つまたは複数の異なるレベル（例えば、音素レベルと同様の）を、特定のレベルに焦点を合わせるためのトレードオフとして、階層的言語構造２００から削除することができる。これらのアプローチでは、アテンション機構３４０は、１つまたは複数の特定の言語層に焦点を合わせるように適宜適合されてもよい。

【００５３】

図４は、テキスト発話３１０の韻律表現３０２を予測する方法４００のための動作の例示的な配置のフローチャートである。方法４００は、図１～図３Ｃを参照して説明することができる。図１のコンピュータシステム１１０上にあるメモリハードウェア１２４は、データ処理ハードウェア１１２によって実行されると、データ処理ハードウェア１１２に方法４００のための動作を実行させる命令を格納していてよい。動作４０２において、方法４００は、テキスト発話３１０を受信することを含む。テキスト発話３１０は、少なくとも１つの単語２４０を有し、各単語２４０は、少なくとも１つの音節２３０を有し、各音節２３０は、少なくとも１つの音素２２０を有する。動作４０４において、方法４００は、テキスト発話３１０のための発話埋め込み２０４を選択することを含む。発話埋め込み２０４は、意図された韻律を表す。本明細書で使用される場合、選択された発話埋め込み２０４は、テキスト発話３１０から意図された韻律を有する合成音声１２２を生成するためにＴＴＳシステム１２０が使用するための、テキスト発話３１０の韻律表現３０２を予測するために使用される。発話埋め込み２０４は、固定長の数値ベクトルによって表すことができる。数値ベクトルは、「２５６」に等しい値を含み得る。テキスト発話３１０に対して発話埋め込み２０４を選択するために、データ処理ハードウェア１１２は、最初にデータストレージ１３０に問い合わせて、テキスト発話３１０に厳密に一致するトランスクリプト２０６を有する発話埋め込み２０４を見つけ、次に、発話埋め込み２０４を選択して、所与のテキスト発話３１０の韻律表現３０２を予測することができる。いくつかの例では、固定長の発話埋め込み２０４は、ターゲット韻律の特定のセマンティクスおよび語用論を表す可能性が高い埋め込み２０４の潜在空間内の特定の点を選ぶことによって選択される。他の例では、潜在空間は、テキスト発話３１０の意図された韻律を表すために、ランダムな発話埋め込み２０４を選択するためにサンプリングされる。さらに別の例では、データ処理ハードウェア１１２は、テキスト発話３１０の言語的特徴に対する最も可能性の高い韻律を表すために、厳密に一致するトランスクリプト２０６を有する発話埋め込み２０４の平均を選択することによって、潜在空間を多次元ユニットガウシアンとしてモデル化する。

【００５４】

動作４０６において、各音節２３０について、選択された発話埋め込み２０４を使用して、方法４００は、音節２３０の各音素２２０に対する言語的特徴２２２へのアテンショ

ン機構３４０によるアテンションに基づいて、音節２３０の韻律音節埋め込み２３２，２３４を復号化することによって、音節２３０の持続時間２３８を予測することを含む。動作４０８において、各音節２３０について、選択された発話埋め込み２０４を使用して、方法４００は、音節２３０に対して予測された持続時間２３８に基づいて、複数の固定長予測フレーム２６０を生成することを含む。

【００５５】

図５は、本文書に記載されているシステム（例えば、階層構造２００、オートエンコーダ３００、および／またはアテンション機構３４０）および方法（例えば、方法４００）を実装するために使用することができる例示的なコンピューティングデバイス５００の概略図である。コンピューティングデバイス５００は、ラップトップ、デスクトップ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、メインフレーム、および他の適切なコンピュータなどの様々な形態のデジタルコンピュータを表すことが意図されている。ここに示されているコンポーネント、それらの接続および関係、並びにそれらの機能は、単なる例示を意味するものであり、この文書で説明および／または主張されている発明の実施を制限することを意味するものではない。

【００５６】

コンピューティングデバイス５００は、プロセッサ５１０（例えば、データ処理ハードウェア）と、メモリ５２０（例えば、メモリハードウェア）と、ストレージデバイス５３０と、メモリ５２０および高速拡張ポート５５０に接続する高速インタフェース／コントローラ５４０と、低速バス５７０およびストレージデバイス５３０に接続する低速インタフェース／コントローラ５６０とを含む。コンポーネント５１０，５２０，５３０，５４０，５５０，および５６０の各々は、様々なバスを使用して相互接続されており、共通のマザーボードに、または必要に応じて他の方法で取り付けることができる。プロセッサ５１０は、高速インタフェース５４０に結合されたディスプレイ５８０などの外部入力／出力デバイスにグラフィカルユーザインタフェース（ＧＵＩ）のためのグラフィカル情報を表示するために、メモリ５２０またはストレージデバイス５３０に格納された命令を含む、コンピューティングデバイス５００内で実行するための命令を処理することができる。他の実装形態では、複数のメモリおよびメモリのタイプとともに、必要に応じて、複数のプロセッサおよび／または複数のバスを使用することができる。また、複数のコンピューティングデバイス５００が接続されてもよく、各デバイスは、（例えば、サーババンク、ブレードサーバのグループ、またはマルチプロセッサシステムとして）必要な動作の一部を提供する。

【００５７】

メモリ５２０は、コンピューティングデバイス５００内に非一時的に情報を格納する。メモリ５２０は、コンピュータ可読媒体、揮発性メモリユニット（複数可）、または不揮発性メモリユニット（複数可）であってよい。非一時的メモリ５２０は、コンピューティングデバイス５００によって使用するために一時的または永続的にプログラム（例えば、命令のシーケンス）またはデータ（例えば、プログラム状態情報）を格納するために使用される物理デバイスであってよい。不揮発性メモリの例には、フラッシュメモリおよび読み取り専用メモリ（ＲＯＭ）／プログラム可能読み取り専用メモリ（ＰＲＯＭ）／消去可能プログラム可能読み取り専用メモリ（ＥＰＲＯＭ）／電子的に消去可能プログラム可能読み取り専用メモリ（ＥＥＰＲＯＭ）（例えば、ブートプログラムなどのファームウェアに通常使用される）が含まれるが、これらに限定されない。揮発性メモリの例には、ランダムアクセスメモリ（ＲＡＭ）、ダイナミックランダムアクセスメモリ（ＤＲＡＭ）、スタティックランダムアクセスメモリ（ＳＲＡＭ）、相変化メモリ（ＰＣＭ）、およびディスクまたはテープが含まれるが、これらに限定されない。

【００５８】

ストレージデバイス５３０は、コンピューティングデバイス５００に大容量ストレージデバイスを提供することができる。いくつかの実装形態では、ストレージデバイス５３０は、コンピュータ可読媒体である。様々な異なる実施形態では、ストレージデバイス５３

０は、フロッピーディスク（登録商標）デバイス、ハードディスクデバイス、光ディスクデバイス、またはテープデバイス、フラッシュメモリまたは他の類似のソリッドステートストレージデバイス、またはストレージエリアネットワークまたは他の構成におけるデバイスを含むデバイスのアレイであってよい。追加の実装形態では、コンピュータプログラム製品は、情報担体に有体的に具現化される。コンピュータプログラム製品は、実行されると、上述したような１つまたは複数の方法を実行する命令を含む。情報担体は、メモリ５２０、ストレージデバイス５３０、またはプロセッサ５１０上のメモリなどの、コンピュータまたは機械で読み取り可能な媒体である。

【００５９】

高速コントローラ５４０は、コンピューティングデバイス５００のための帯域幅集約的な動作を管理し、一方、低速コントローラ５６０は、より少ない帯域幅を消費する動作を管理する。このような職務の割り当ては例示に過ぎない。いくつかの実装形態では、高速コントローラ５４０は、メモリ５２０、ディスプレイ５８０（例えば、グラフィックプロセッサまたはアクセラレータを介して）に結合され、および様々な拡張カード（図示せず）を受け入れることができる高速拡張ポート５５０に結合される。いくつかの実装形態では、低速コントローラ５６０は、ストレージデバイス５３０および低速拡張ポート５９０に結合されている。様々な通信ポート（例えば、ＵＳＢ、ブルートゥース（登録商標）、イーサネット（登録商標）、ワイヤレスイーサネット）を含み得る低速拡張ポート５９０は、キーボード、ポインティングデバイス、スキャナなどの１つまたは複数の入力／出力デバイスに、またはスイッチやルーターなどのネットワーキングデバイスに、例えば、ネットワークアダプタを介して結合され得る。

【００６０】

コンピューティングデバイス５００は、図に示されるように、複数の異なる形態で実装することができる。例えば、それは、標準サーバ５００ａとして、またはそのようなサーバ５００ａのグループに複数回、ラップトップコンピュータ５００ｂとして、またはラックサーバシステム５００ｃの一部として実装することができる。

【００６１】

本明細書に記載のシステムおよび技術の様々な実装形態は、デジタル電子および／または光回路、集積回路、特別に設計されたＡＳＩＣ（特定用途向け集積回路）、コンピュータハードウェア、ファームウェア、ソフトウェア、および／またはそれらの組み合わせで実現することができる。これらの様々な実装形態は、ストレージシステム、少なくとも１つの入力デバイス、および少なくとも１つの出力デバイスからデータおよび命令を受信し、それらへデータおよび命令を送信するために結合された、専用または汎用であり得る少なくとも１つのプログラム可能なプロセッサを含むプログラム可能なシステム上で実行可能および／または解釈可能な１つまたは複数のコンピュータプログラムにおける実装形態を含むことができる。

【００６２】

これらのコンピュータプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーション、またはコードとしても知られる）は、プログラム可能なプロセッサのマシン命令を含み、高レベルの手続き型および／またはオブジェクト指向のプログラミング言語、および／またはアセンブリ／マシン言語で実装することができる。本明細書で使用される場合、「機械可読媒体」および「コンピュータ可読媒体」という用語は、機械命令および／またはデータをプログラム可能なプロセッサに提供するために使用される任意のコンピュータプログラム製品、非一時的なコンピュータ可読媒体、装置および／またはデバイス（例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジックデバイス（ＰＬＤ））を指し、機械命令を機械可読信号として受信する機械可読媒体を含む。「機械可読信号」という用語は、プログラム可能なプロセッサに機械命令および／またはデータを提供するために使用される任意の信号を指す。

【００６３】

本明細書に記載のプロセスおよび論理フローは、１つまたは複数のコンピュータプログ

ラムを実行して、入力データを操作し、出力を生成することによって機能を実行する１つまたは複数のプログラム可能なプロセッサによって実行することができる。プロセスおよびロジックフローは、例えば、ＦＰＧＡ（フィールドプログラマブルゲートアレイ）またはＡＳＩＣ（特定用途向け集積回路）などの特定用途のロジック回路によっても実行できる。コンピュータプログラムの実行に適したプロセッサは、例として、汎用マイクロプロセッサおよび専用マイクロプロセッサの両方、および任意の種類のデジタルコンピュータの任意の１つまたは複数のプロセッサを含む。一般に、プロセッサは、読み取り専用メモリまたはランダムアクセスメモリ、あるいはその両方から命令とデータを受け取る。コンピュータの重要な素子は、命令を実行するためのプロセッサと、命令およびデータを格納するための１つまたは複数のメモリデバイスである。一般に、コンピュータは、データを格納するための１つまたは複数の大容量ストレージデバイス、例えば、磁気、光磁気ディスク、または光ディスクを含むか、またはそれらからデータを受信し、または転送し、あるいはその両方のために動作可能に結合される。しかしながら、コンピュータがそのようなデバイスを有する必要はない。コンピュータプログラム命令およびデータを格納するのに適したコンピュータ可読媒体は、あらゆる形態の不揮発性メモリ、媒体およびメモリデバイスを含み、それらは、例えば、半導体メモリデバイス、例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭ、およびフラッシュメモリデバイス、磁気ディスク、例えば、内蔵ハードディスクまたはリムーバブルディスク、光磁気ディスク、およびＣＤ　ＲＯＭおよびＤＶＤ－ＲＯＭディスクを含む。プロセッサおよびメモリは、特定用途の論理回路によって補完されてよく、または特定用途の論理回路に組み込まれてもよい。

【００６４】

ユーザとのインタラクションを提供するために、本開示の１つまたは複数の態様は、ユーザに情報を表示するためのディスプレイデバイス、例えば、ＣＲＴ（陰極線管）、ＬＤＣ（液晶ディスプレイ）モニタ、またはタッチスクリーンを有するコンピュータ上に実装されてもよく、コンピュータは、任意選択で、キーボードおよびポインティングデバイス、例えば、マウスまたはトラックボールを有し、これによって、ユーザがコンピュータに入力を提供できる。他の種類のデバイスを使用して、ユーザとのインタラクションを提供することもでき、例えば、ユーザに提供されるフィードバックは、例えば、視覚的フィードバック、聴覚的フィードバック、または触覚的フィードバックなど、任意の形態の感覚的フィードバックであってよく、ユーザからの入力は、音響、音声、または触覚入力を含む任意の形式で受け取ることができる。加えて、コンピュータは、ユーザによって使用されるデバイスとの間でドキュメントを送受信することにより、ユーザと対話（ｉｎｔｅｒａｃｔ）でき、これは、例えば、Ｗｅｂブラウザから受信された要求に応答して、ユーザのクライアントデバイス上のＷｅｂブラウザにＷｅｂページを送信することによって行われる。

【００６５】

複数の実装形態が説明された。それにもかかわらず、本開示の技術思想および範囲から逸脱することなく、様々な修正を行うことができることが理解されるであろう。したがって、他の実装形態は、以下の特許請求の範囲内にある。

【図面】

【図１】

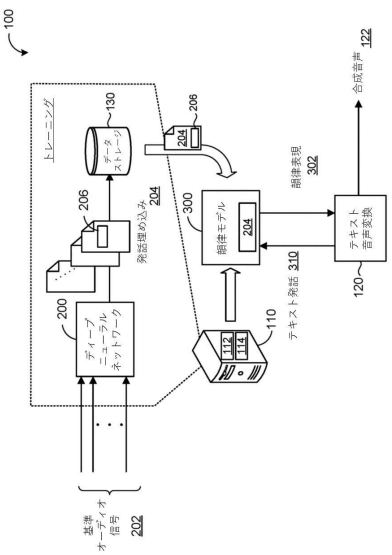

FIG. 1

【図２Ａ】

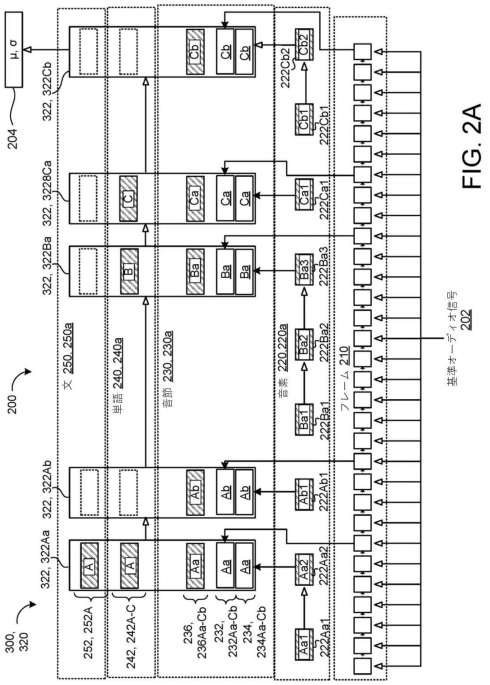

FIG. 2A

【図２Ｂ】

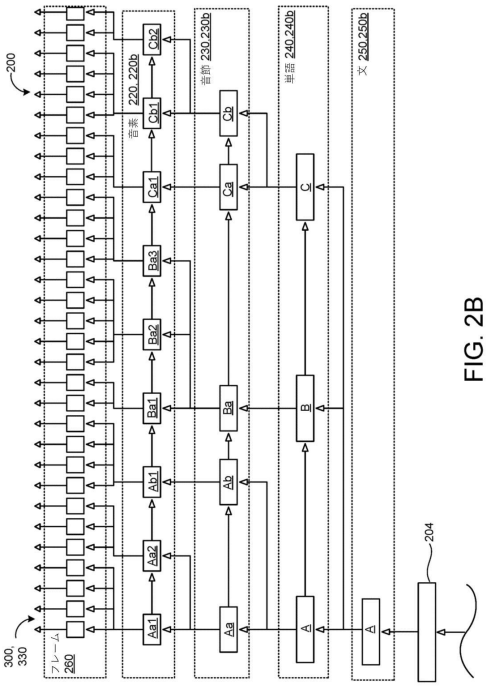

FIG. 2B

【図２Ｃ】

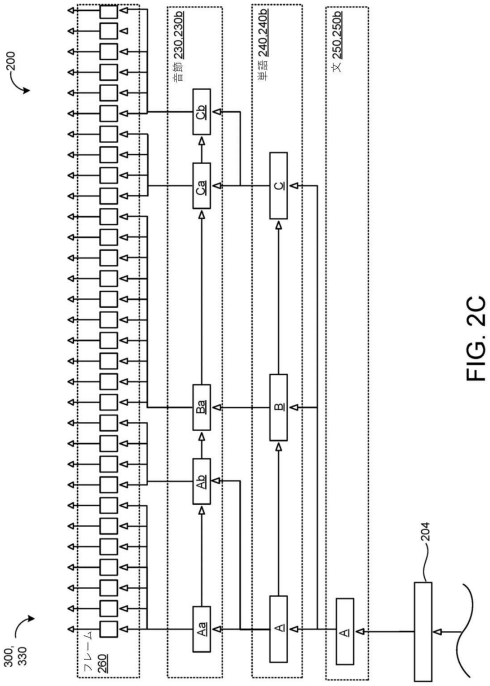

FIG. 2C

【図３Ａ】

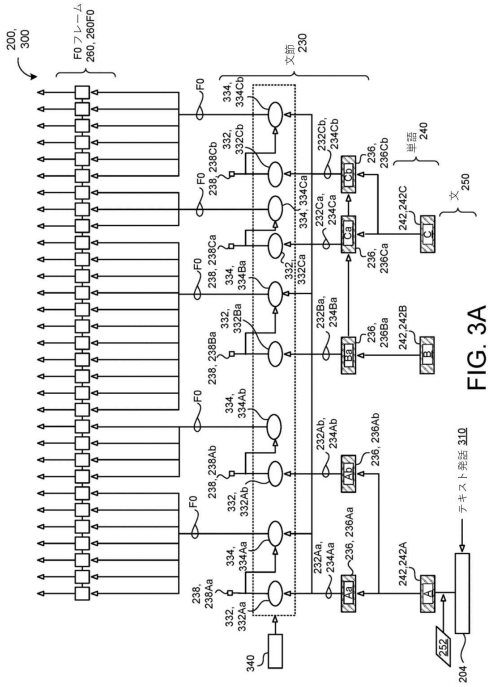

FIG. 3A

【図３Ｂ】

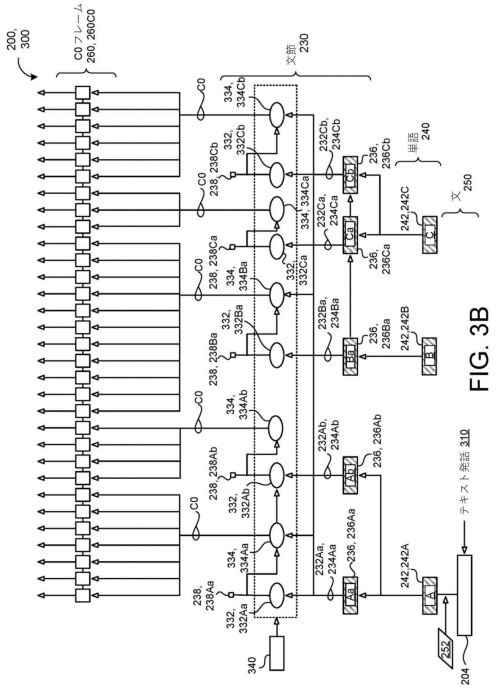

FIG. 3B

【図３Ｃ】

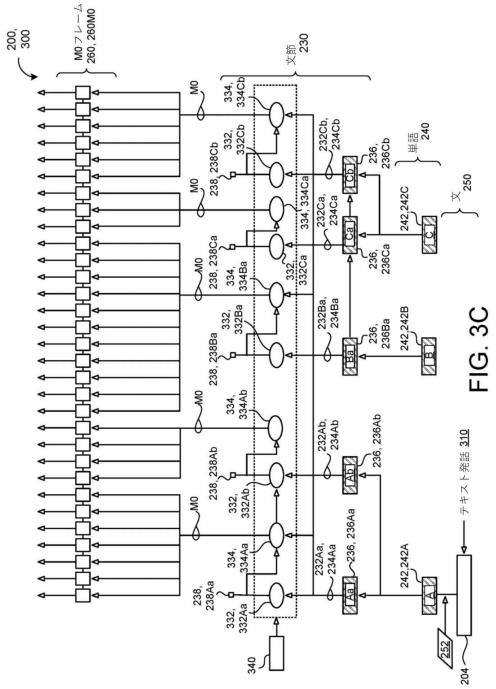

FIG. 3C

【図３Ｄ】

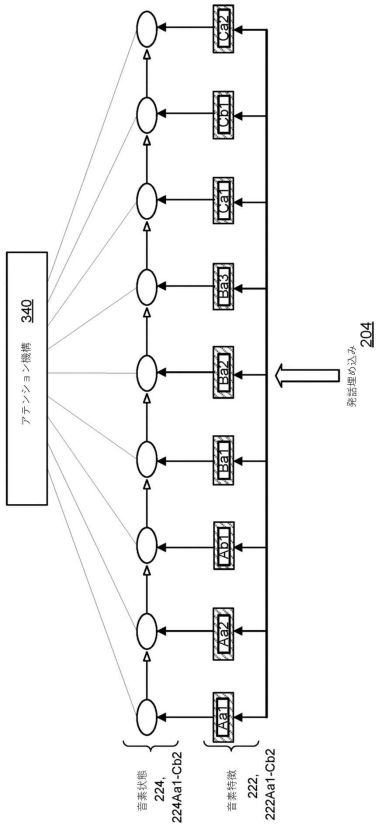

FIG. 3D

【図４】

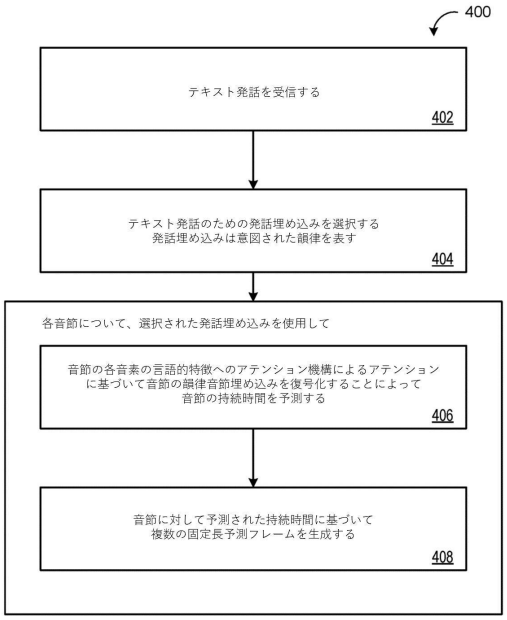

FIG. 4

【図５】

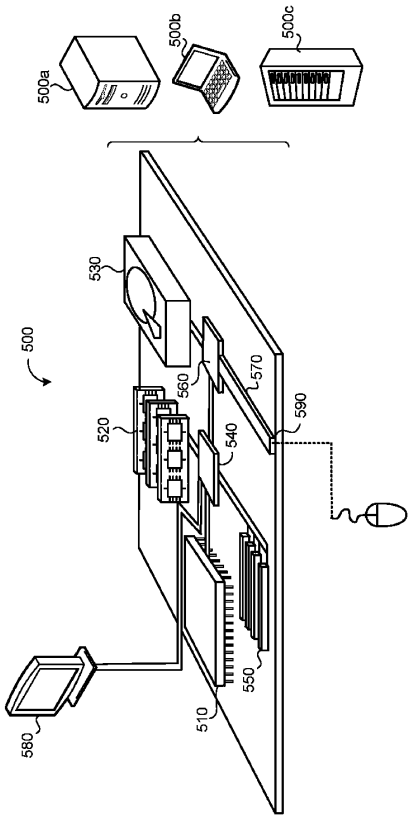

FIG. 5

フロントページの続き

(72)発明者　チャン、チュン－アン
アメリカ合衆国 ９４０４３ カリフォルニア州 マウンテン ビュー アンフィシアター パークウェイ １６００
(72)発明者　ワン、ヴィンセント
アメリカ合衆国 ９４０４３ カリフォルニア州 マウンテン ビュー アンフィシアター パークウェイ １６００
審査官　山下 剛史
(56)参考文献　国際公開第２０１７／１６８８７０（WO，Ａ１）
国際公開第２０１９／２１７０３５（WO，Ａ１）
田中宏他，ＶＡＥ－ＳＰＡＣＥ：音声Ｆ０パターンの深層生成モデル，日本音響学会２０１８年春季研究発表会講演論文集［ＣＤ－ＲＯＭ］，2018年03月，pp.229-230
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ １３／００－１３／１０，２５／３０
ＩＥＥＥ Ｘｐｌｏｒｅ